APPENDIX (A)

Task No. 7

The Pennsylvania State College
Mineral Industries Experiment Station
Division of Meteorology

Progress Report No. 1
on

INVESTIGATION OF THE OPTIMUM ANALYTICAL TECHNIQUES FOR WEATHER ANALYSIS AND FORECASTING

30 September 1952

Prepared for
Bureau of Aeronautics Project AROWA (TED-UNL-MA-501)
under
Contract No. N189s-86996

THE PENNSYLVANIA STATE COLLEGE

MINERAL INDUSTRIES EXPERIMENT STATION

DIVISION OF METEOROLOGY

Progress Report No. 1

Contract No. N189s-86996

Submitted to the Bureau of Aeronautics Project AROWA, Naval Air Station, Norfolk, Virginia. The work reported herein is of a preliminary nature and the results are not necessarily in final form.

INVESTIGATION OF THE OPTIMUM ANALYTICAL TECHNIQUES FOR WEATHER ANALYSIS AND FORECASTING

Prepared by:

V. Moyer, Research Associate in Meteorology

Approved by:

H. Neuberger, Chief, Division of Meteorology

30 September 1952

TABLE OF CONTENTS

LIST OF FIGURES

1.0 Abstract

Research into one of the possible approaches to the investigation of optimum analytical and forecasting techniques in weather analysis and forecasting was initiated on 1 July 1952. In effect, this approach consists of withholding from the meteorologist certain combinations of synoptic data, yet requiring him to forecast all the usual weather parameters including those withheld.

During the time that the necessary charts were being prepared for the main project, a pilot test was conducted for the purpose of determining the characteristics of the experimental design. The results of the pilot test for sea-level and 500-mb data indicate (1) that an effective deterioration of forecast scores is noticeable only for severely reduced information; (2) that variation of the forecast verification schemes merely shifts the error level of forecasts without affecting the score trends; and (3) that there exists an optimum scheme of reduction of data furnished to the meteorologist which results in a considerable saving of time required for spotting, analyzing, and forecasting, without materially affecting forecast scores.

In addition, an experiment was conducted with a class of meteorology trainees; one section of this class was required to forecast from previously prepared prognostic charts that actually represented correct analyses of the subsequent maps; another section of the class was required to construct their own prognostic charts. The results of this experiment cannot be considered conclusive at this time.

2.0 Personnel and Administration

2.1 The following persons were employed on the project during the period in question: Dr. H. Neuberger, Professor of Meteorology and Project Supervisor, approximately 1/3-time without cost to the project; Mr. V. Moyer, Research Associate and Assistant Project Supervisor, full-time; Mr. I. Van der Hoven, Research Assistant, full-time after 1 September 1952; and Mr. David L. Jones, Graduate Assistant, 3/4-time. Mr. Howard Simmons, Supervisory Meteorologist, U. S. Weather Bureau, Evansville, Indiana, is participating in the analysis and forecasting of the final test data without cost to the project. Dr. Hans Panofsky, Associate Professor; Dr. Charles L. Hosler, Jr., Assistant Professor, and Mr. Donald G. Yerg, Instructor

in Meteorology, assisted from time to time in the capacity of consultants without cost to the project.

2.2 The following special students and upper classmen served as technical assistants on a part-time, hourly basis:

Mr. James P. Anderson

Mr. Floyd C. Elder

Mr. William Holtzman

Mr. Lowell Krawitz

Mr. Harry R. Mansfield

Mr. Evan J. Tibbott

Mr. Prodipto Roy

Mr. Robert B. Wassall

2.3 The secretarial work was performed, until 1 September 1952, by Mrs. Kay Fisher at no expense to the project. After 1 September, Mrs. Mary Wagner was engaged as full-time project secretary; she was assisted from time to time by Mrs. Peggy Rase, without cost to the project.

3.0 Conferences

3.1 Between 23 and 27 September 1952, Dr. Neuberger and Mr. Moyer undertook a trip to Cambridge, Massachusetts, and New York City for the purpose of conferring with Professor G. P. Wadsworth, Dr. J. G. Bryan, and Mr. William Paulsen, at Massachusetts Institute of Technology, and with Professor James E. Miller, at New York University, on the problem of the accuracy of short-range weather forecasts, the limitations of weather verification systems, and the philosophy of non-isobaric analysis. While in Cambridge, the above project members also visited the Air Force Cambridge Research Center of the Geophysical Research Directorate where they met with Drs. Richard A. Craig and William K. Widger, and Messrs. Irving I. Gringorten and Ivor Lund in a discussion of forecast verification. The results of the conferences can be summarized as follows: (1) The handicap of reduced information, particularly the omission of barometric pressure from surface weather maps, seems to be more psychological than physical. The amount of information currently presented is based on tradition rather than on

scientifically well considered needs. As a result, the experienced forecaster revolts at the forced task of forecasting from anything but the complete map to which he is accustomed; (2) two general classes of forecast verification systems can be considered: the multiple error point system in which one or more error points are scored depending on the degree by which the forecast item was missed; or the binary system, in which an item forecast is either right or wrong. The latter system appears to be preferable, notably from the point of view of utility and of minimizing bias. This supports our original position in this matter.

3.2 Mr. David Jones visited the National Weather Records Center, Asheville, North Carolina, on 29 August 1952, while on annual leave. Through the active assistance of Mr. Lesley Smith, Supervisor, he was able to obtain the scattered data that were missing from the Ozalid WBAN Analysis Center maps which were used as source of synoptic information.

3.3 LCDR Donald R. Jones, AROWA Project Officer, visited the project in State College on 5 September 1952 for the purpose of discussing procedures. His visit was followed by telephone conversations both with him and with LCDR William J. Kotsch about misunderstandings over the over-all purpose and procedures of the experiment, as well as contract details. Considerable correspondence was required, subsequently, in connection with ordering supplies and materials.

4.0 Investigations being undertaken

Prior to the 1 July 1952 start of the work, complete plans had been formulated to put the project into operation on the starting date without delay. An office, Room 318, was set aside in the Mineral Industries Building for the sole use of project personnel. In addition to the usual equipment and the necessary minor supplies, it was furnished with six drafting tables and a portable light table to expedite plotting and analysis of maps. This office serves as permanent quarters for Messrs. Moyer, Van der Hoven, and Da... Jones, and provides working space for part-time technical assistants.

The preliminary planning included tentative decisions on the stations for which forecasts were to be made, forecast items, verification systems and tolerances, and the specific meteorological situations to be tested. Dr. Hosler assisted Dr. Neuberger in choosing representative winter and summer synoptic situations for testing. The other project personnel

were not consulted during this choice, since it was desired to keep them in ignorance of even the broad features of the synoptic maps that they would later be required to analyze and prognosticate.

From the multitude of possible "schemes" of synoptic weather data presentation (factorial 12), five logical combinations were chosen for testing in a pilot test prior to the start of the final analysis of the two selected situations. Dr. Panofsky and Mr. Yerg were consulted in the choice of these schemes because of the statistical implications involved in the future analysis of the results.

Also, Mr. Moyer developed a new base map, of convenient size and scale, that would include the forecast area as well as an adequate "influence zone" to the west of this area (Enclosure I). This map was printed by offset process in sufficient quantity for the duration of the project.

Because of the magnitude of the program with consideration of time consumption in spotting, analyzing, and forecasting from the test maps, the forecast stations were restricted to the area between 25° N and 65° N latitude and 30° W and 98° W longitude. However, the influence zone was extended to 110° W longitude. Within this area, the forecast stations selected were:

(1) Weather ship "Alpha," 4YA

(2) Weather ship "Beta," 4YB

(3) Moosonee, Ontario, 836

(4) Weather ship "Coca," 4YC

(5) International Falls, Minnesota, 747

(6) Caribou, Maine, 712

(7) Torbay, Newfoundland, 801

or

Pepperrell AFB, Newfoundland, 198

(8) Chicago (Joliet), Illinois, 534 (JOT)

(9) Buffalo, New York, 528

(10) Nantucket, Massachusetts, 506

(11) Weather ship "Delta," 4YD

(12) Kansas City, Missouri (Fort Leavenworth, Kansas), 446 (FLV)

(13) Nashville, Tennessee, 327

(14) Hatteras, North Carolina, 304

(15) Shreveport (Barksdale AFB), Louisiana, 248 (BAD)

(16) Pensacola, Florida, 222

(17) Kindley AFB, Bermuda, 016

(18) Weather ship "Echo," 4YE

(19) Brownsville, Texas , 250

(20) Miami, Florida, 202

The stations listed in parentheses are the upper air sounding stations nearest the indicated surface stations.

Although it had originally been anticipated that I.B.M. methods would be used to analyze the results of the tests, these procedures were found, after consultation with College experts, to be too inflexible for our purpose; therefore, McBee Keysort cards were designed to facilitate subsequent analysis (Enclosure II)

4.1 Pilot test. During the early weeks of July all project members were occupied in plotting the data for a pilot test which was designed to test the experimental schemes of data presentation and analysis. Teletype data were used for this test, there being six maps analyzed for each scheme. The situation analyzed covered the period from 1235Z of 1 July 1952 through 0035Z of 4 July 1952, the continuity followed, thus, consisting of 12 hours.

For purposes of plotting ease, the synoptic code was rearranged as follows:

(1) Land Stations: iiidd ffNww $C_L N_h h_c C_M C_H$
$4VVT_dT_d$ TTPPP $appWR_t$ $7RRD_cs$

(2) Sea Stations: ØØØ LLL dd ffNww $C_L N_h h_c C_M C_H$
$4VVT_dT_d$ TTPPP $appWR_t$ D_sV_s

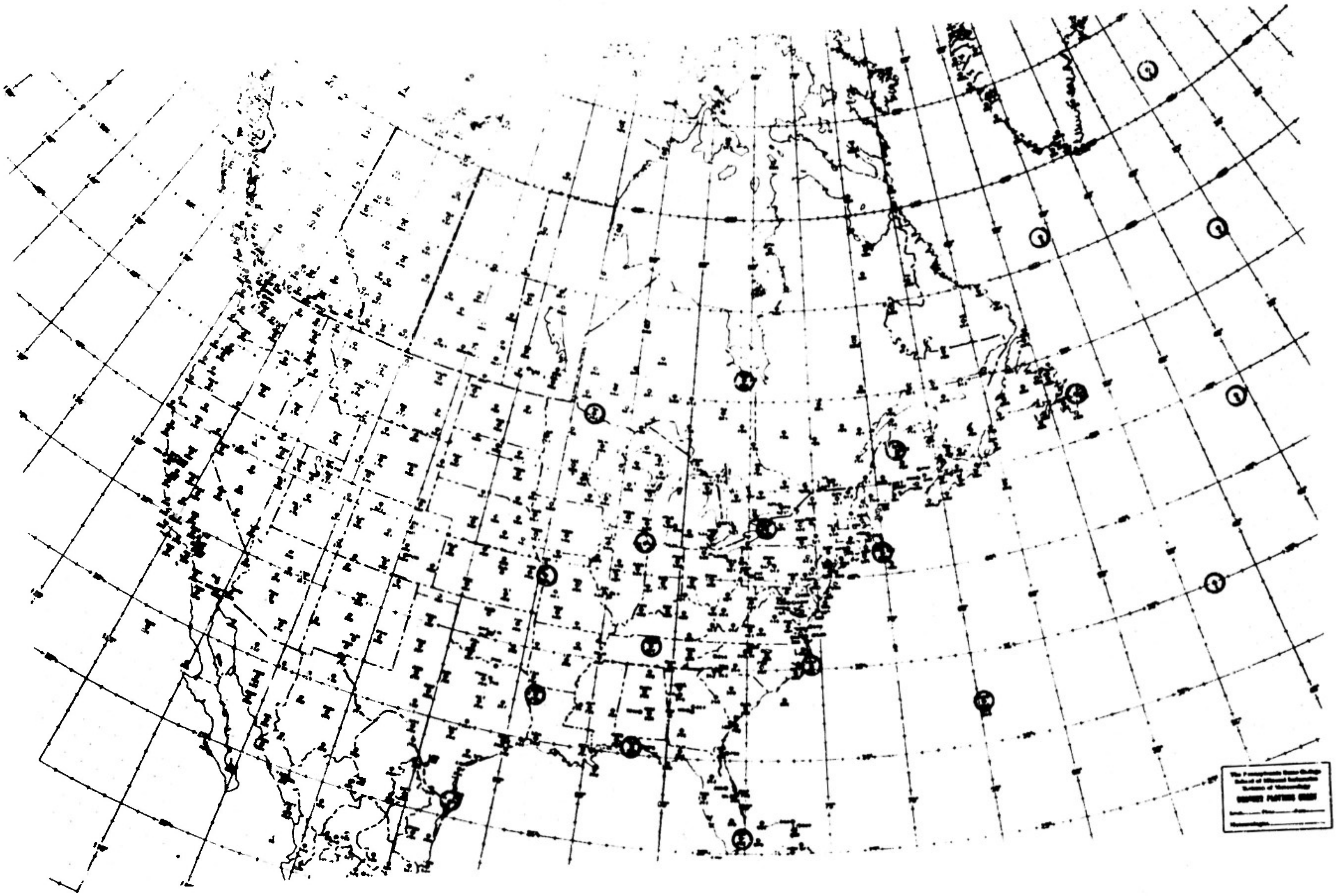


Fig. 1. Forecast stations for final investigation

Also, mimeographed forms were used which contained the reporting stations in a given order to facilitate spotting with Ditto ink: i.e., the stations were listed from within latitude bands running from west to east, starting at the north of the map, such that the spotter could avoid smudging data already spotted and not yet dry.

The following schemes were chosen for providing synoptic data:

E.	iii//	///ww	/ / / / /	4/// /	TTPPP	/////	//// /
D.	iii//	//Nww	C_L/ / $C_M C_H$	4/// /	TTPPP	app//	//// /
C.	iiidd	ffNww	C_L// $C_M C_H$	4VV$T_d T_d$	TT///	appWR_t7RR//	
B.	iii//	//Nww	C_L/ / $C_M C_H$	4VV$T_d T_d$	TTPPP	appWR_t7RR//	
A.	iiidd	ffNww	$C_L N_h h_c C_M C_H$	4VV$T_d T_d$	TTPPP	appWR_t7RRD_cs	

All six maps of each scheme were analyzed and prognosticated before work with the next scheme was started. The schemes were analyzed in reverse order viz., scheme "E", scheme "D", etc. While the technical assistants were occupied in transcribing data for the main test, from the Ozalid copies of WBAN-1 sea-level and 500-mb analyses of the WBAN Analysis Center supplied to us by LCDR Donald R. Jones, Messrs. Moyer and David Jones conducted the pilot experiment. Neither of these men compared analyses during the test, nor were their forecasts verified until all had been submitted. Each man drew from his personal experience in analysis and forecasting and used whatever device was possible within the limitations of each given scheme. No attempt was made to force either man to use any of the so-called standard methods of prognosis. That is to say, if one of the forecasters chose, for a particular map, to be completely subjective and to rely entirely upon intuition, he was at liberty to do so; on the other hand, if he attempted to apply objective forecasting rules and methods in his prognosis again he was free to do so.

Because of our desire to accomplish this test within a minimum of time, only incomplete upper air charts were at first available; they consisted of 500-mb charts of irregular continuity, drawn by students in their synoptic laboratory course. However, after it became apparent that the pilot test could be

completed well within our self-imposed deadline, a similar experiment was conducted with 500-mb data for the same period, with charts specially prepared for this purpose. The five schemes of data presentation at this level were:

E. iiidd   ff///   /// /

D. iii//   //hhh   /// /

C. iiidd   ff///   TT/ /

B. iii//   //hhh   TT/ /

A. iiidd   ffhhh $TTT_d T_d$

The procedure in this phase of the pilot experiment was precisely the same as that for the sea-level data. All six charts of scheme "E" were analyzed and prognosticated before the forecaster turned his attention to scheme "D", etc. During this phase, the previously-analyzed "A" scheme of sea-level maps was used as supplementary information.

No time limits or deadlines were imposed upon the forecasters during the test run. Since a time study was conducted for this experiment on the same basis as that planned for the final test, it was believed essential to let each individual set his own pace. Also, because the personality of the forecaster certainly enters into the facility with which he arrives at a forecast, the results obtained were not biased because of this factor. Every effort was made to achieve realism during the process of the experiment, except that it was not possible for the forecaster to determine the trend of verification of his forecasts, as would be the case during normal weather station operation. This exception is unfortunate, perhaps, since it did not permit normal adjustments of analyses or forecasts in conformity with developments in the synoptic pattern; on the other hand, this practise resulted in an investigation of minimum operating procedures, since the forecasters were required to work under what we can assume to be the extreme possible handicaps.

To accelerate accomplishment of the goal of the pilot test, only ten of the forecast stations named above were considered. These were:

(1) International Falls

(2) Kansas City (Fort Leavenworth)

(3) Nashville

(4) Hatteras

(5) Miami

(6) Bermuda

(7) Ship "Echo"

(8) Caribou

(9) Ship "Delta"

(10) Ship "Alpha"

Sea-level forecast items included (1) wind direction and speed, (2) sky condition or present weather, (3) ceiling, (4) visibility, (5) special phenomena (fog or thunderstorm), (6) precipitation amount, (7) temperature, and (8) dew point. 500-mb forecast items included (1) height-change, (2) temperature-change, and (3) wind direction and speed. Forecasts were verified on the basis of slight modification of the system suggested in Pulk and Murphy, Workbook for Weather Forecasting, Prentice Hall, New York, 1950, p.43.

4,11 Results of the Pilot Test

The main results of the pilot test for the surface data are as follows:

(1) The maximum difference in score between the two forecasters for a given time and scheme was 22%, whereas the maximum difference between station forecasts for one forecaster was 37%. Table I gives the ranges of the per cent errors of the total scores for each forecaster and for various schemes.

Table I. Ranges of per cent errors of total surface scores for two forecasters and various schemes.

Schemes	Forecaster I	Forecaster II	Both Forecasters
A	16-45	4-41	10-43
B	20-43	4-32	12-34
C	23-42	6-41	15-41
D	16-50	9-38	12-42
E	18-48	10-40	14-42

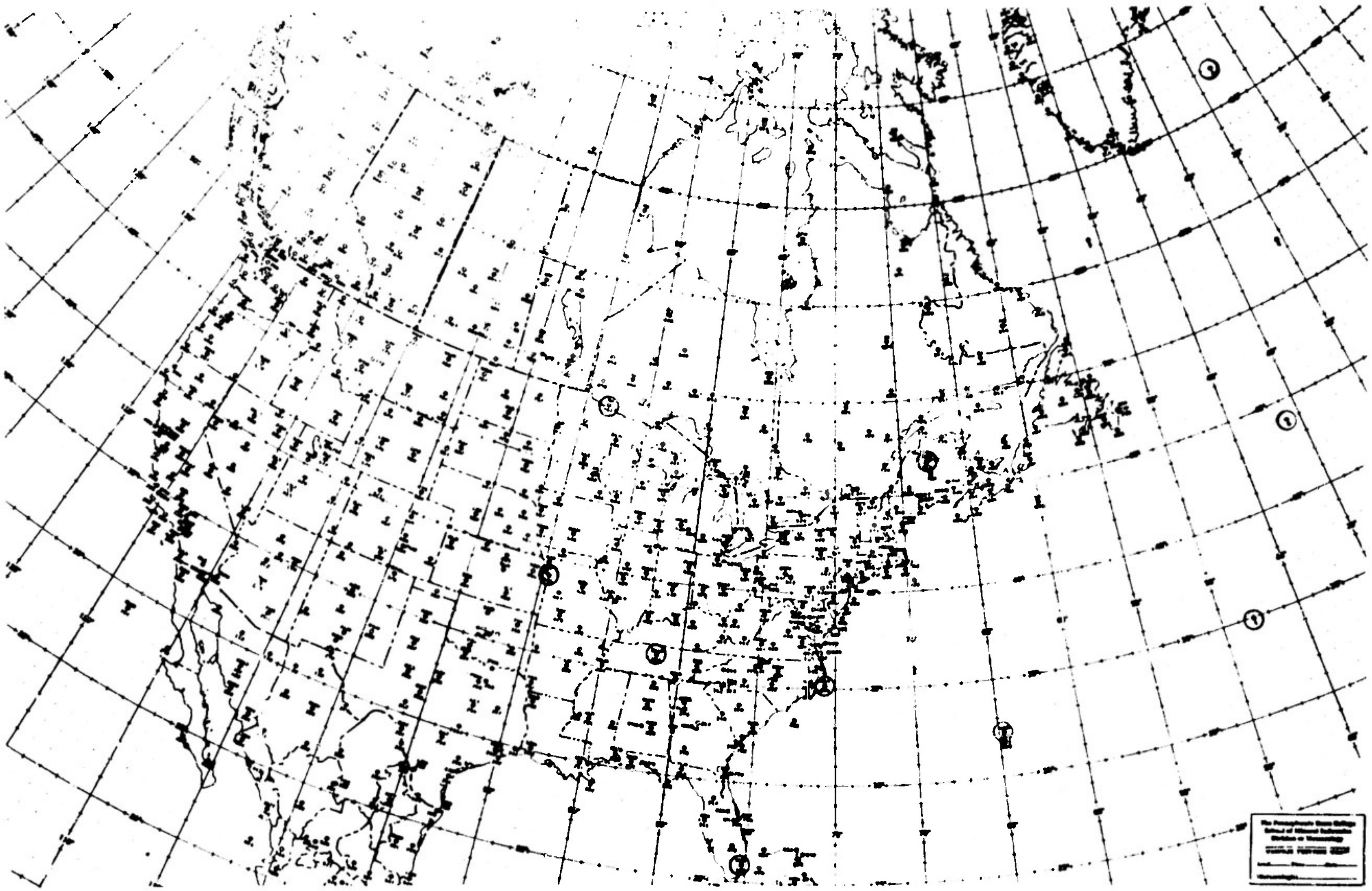


Fig. 2. Forecast stations for pilot test.

(2) The difference in score between 12-and 24-hour forecasts for each forecaster and for both forecasters was negligible as compared to the difference between other parameters such as stations and forecast item. The maximum difference in score for forecaster I was 10% and for forecaster II, 17%. Table II gives the average per cent errors for the various schemes for both 12-and 24-hour forecasts. The number of forecast items per man per scheme was 432 for 12-hour and 440 for 24-hour forecasts. The total number of surface items forecast by each man was 4360.

Table II. Average per cent errors for various schemes and for 12-and 24-hour surface forecasts.

Scheme	Forecaster I		Forecaster II		Both Forecasters	
	12 hr.	24 hr.	12 hr.	24 hr.	12 hr.	24 hr.
A	26	28	24	24	25	26
B	31	28	20	20	26	24
C	32	30	22	28	27	29
D	32	31	28	27	30	29
E	33	34	31	31	32	33

(3) The effect of the scheme on the score for various stations was very irregular, although the scores for 6 of 10 stations were best for scheme B. The forecasts for 3 of the 4 remaining stations were second best for this same scheme. Table III shows the average per cent errors for various stations and schemes, with the best score for each station underlined.

Table III. Average per cent surface forecast errors for various stations and schemes.

Scheme	STATION NUMBER										Average for all stations
	747	712	446	327	304	202	016	4YA	4YD	4YE	
A	32	20	21	28	20	26	15	43	33	10	25
B	34	19	19	27	16	25	28	34	31	12	25
C	33	21	23	26	22	28	29	41	40	15	27
D	29	22	25	32	29	32	32	42	37	12	30
E	41	26	32	30	32	29	37	42	40	14	33

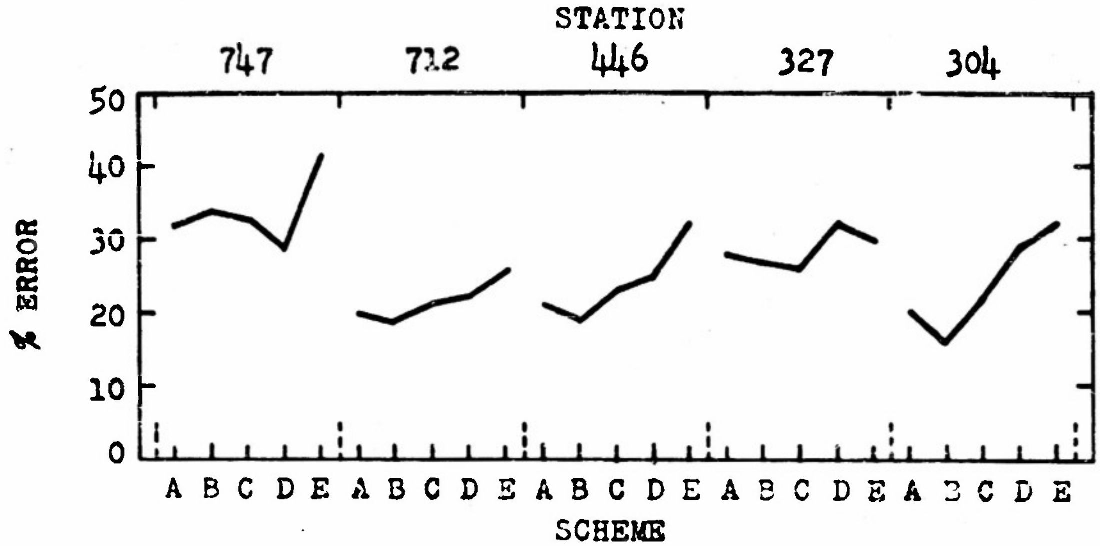
Figure 3 shows graphically the pronounced trend toward score deterioration that exists with reduced information for stations 016, 446, 304, and 712. There is very little trend for stations 327, 202, 4YD, and 4YE, while stations 747 and 4YA show irregular score variations. However, it can be noted that those stations with low error scores show a more significant trend toward score deterioration with reduced information, than those with high error scores.

For a further appraisal of the effect of the various schemes on the forecast scores, the frequencies of four groups of per cent errors were determined for the several schemes. Figure 4 shows this relative frequency distribution. A significant deterioration in score is evident only for schemes D and E where the total frequency of the two highest error groups is 45% and 65%, respectively.

(4) The effect of the schemes on scores for various forecast items is strong only for present weather (WW), temperature (TT), and dew point (T_dT_d), with a slight effect for wind force (FF) and no effect for visibility (VV), wind direction (DD), and rainfall (RR). For cloud heights (h_ch_c), the forecast seems to improve slightly with reduced information; but this result cannot be considered significant. Table IV summarizes the effect of the schemes on the forecast items. The underlined values represent schemes which do not include the parameter listed above the indicated value.

Table IV. Effect of schemes on per cent error score for various forecast items.

Scheme	DD	FF	WW	h_ch_c	VV	RR	TT	T_dT_d	Average
A	19	27	38	35	12	29	17	27	26
B	20	24	37	37	13	26	17	25	25
C	25	26	41	35	11	30	24	33	28
D	23	29	43	25	7	32	29	48	30
E	21	31	48	30	13	30	36	51	32



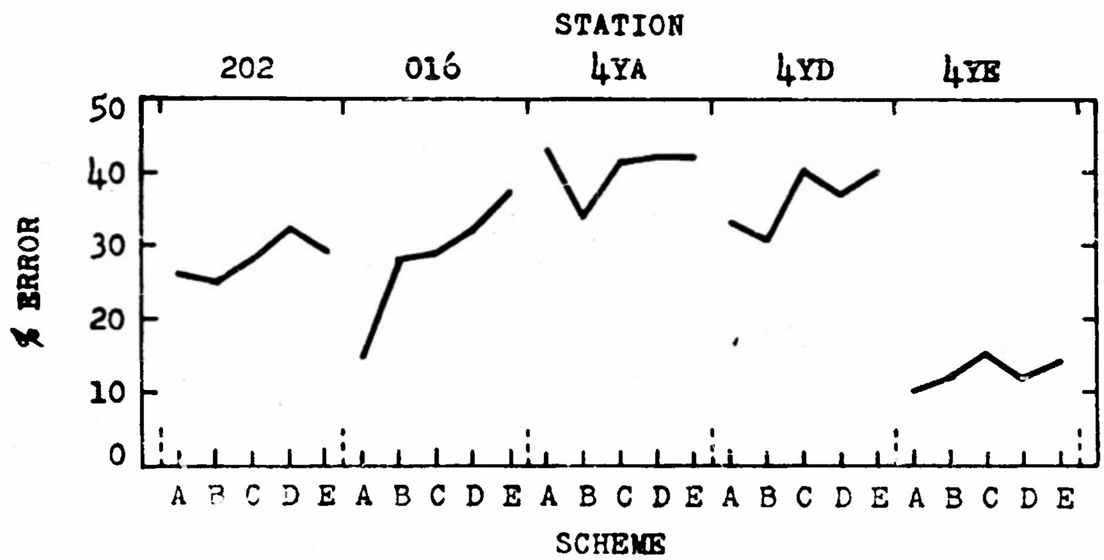


Fig. 3. Average Surface Forecast Errors (%) for Various Stations. Schemes A Through E.

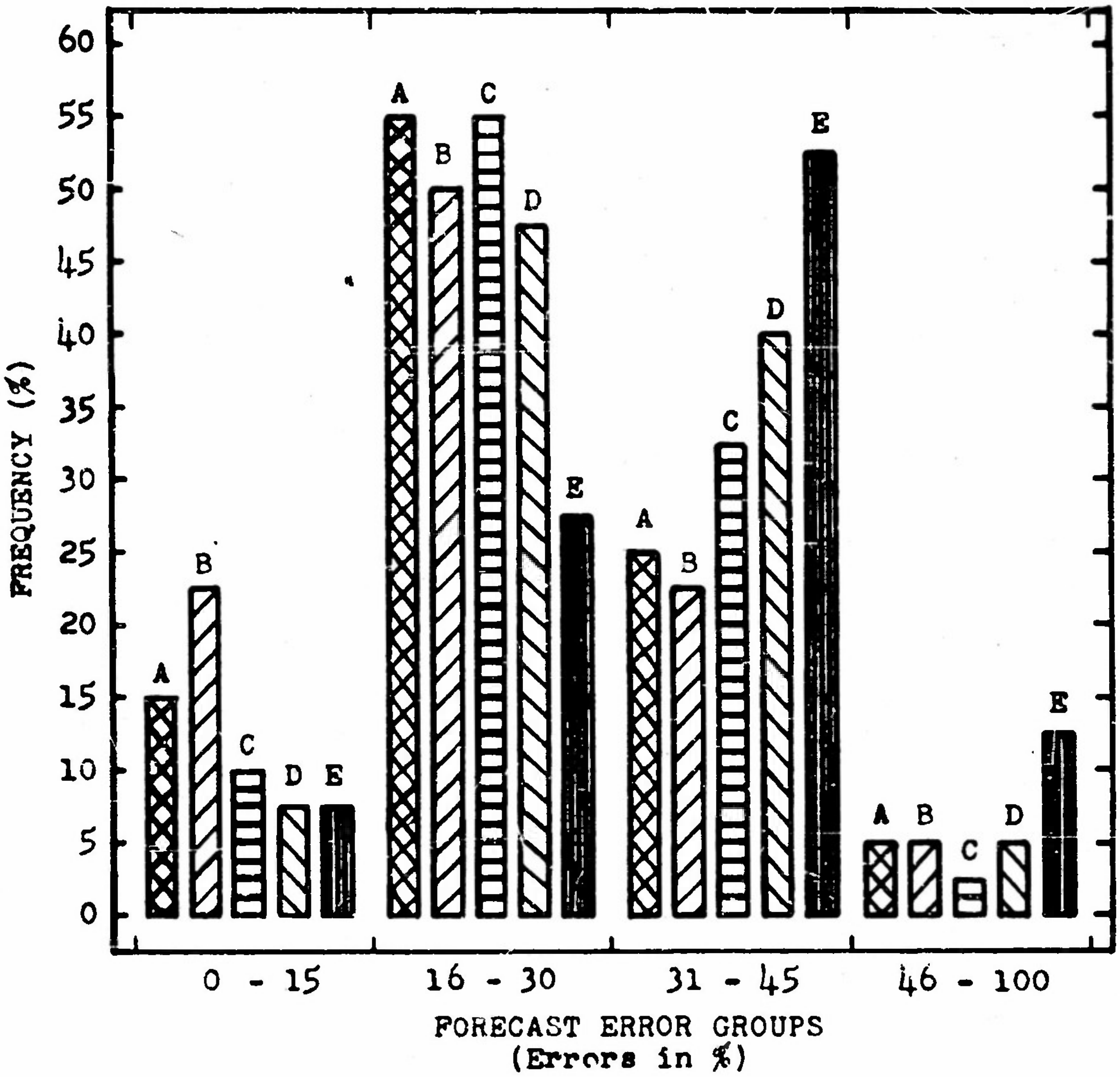


Fig. 4. Frequency Distributions of Surface Forecast Errors (Errors in %). Schemes A Through E.

(5) The effect of the verification system on the results as presented in Table IV was tested by employing a second, slightly more rigorous verification system. Figure 5 clearly shows that a more rigorous verification system increases the error score without materially affecting the trend introduced by the schemes.

From the above results the following conclusions are drawn and discussed:

(1) The scores of the two forecasters and of 12- and 24-hour forecasts can be taken together and treated as one statistical population. This appears to be in agreement with the results presented by G. P. Wadsworth*. Whether or not this will still hold in the main test of the project when more forecasters are employed whose backgrounds are no longer as homogeneous as in the pilot test will have to be investigated by means of the main forecast material.

(2) In general, the trend toward score deterioration with reduced information is more evident for stations with low error scores than for those with high error scores. This may appear to be a trite result, because, e.g., in the absurd case of completely wrong forecasts for scheme A, the score could not possibly deteriorate with reduced information. However, it can be seen from Table III that even the worst score for station 4YA could have deteriorated to a considerable extent.

(3)The effect of weather variability at various stations is apparent from the trends of the scores. This id particularly evident when the scores for station 4YE in the subtropical anticyclone are compared with those of station 4YA which lies in the North Atlantic storm track. This clearly points to the necessity for separate analysis of the forecasts for individual stations during the main project.

(4) The ease of forecasting varies widely for

---

*"Accuracy of Short-Range Forecasting Comparative Evaluations", GRD Contract No. AF19(122)447, G. P. Wadsworth, Project Director, Massachusetts Institute of Technology.

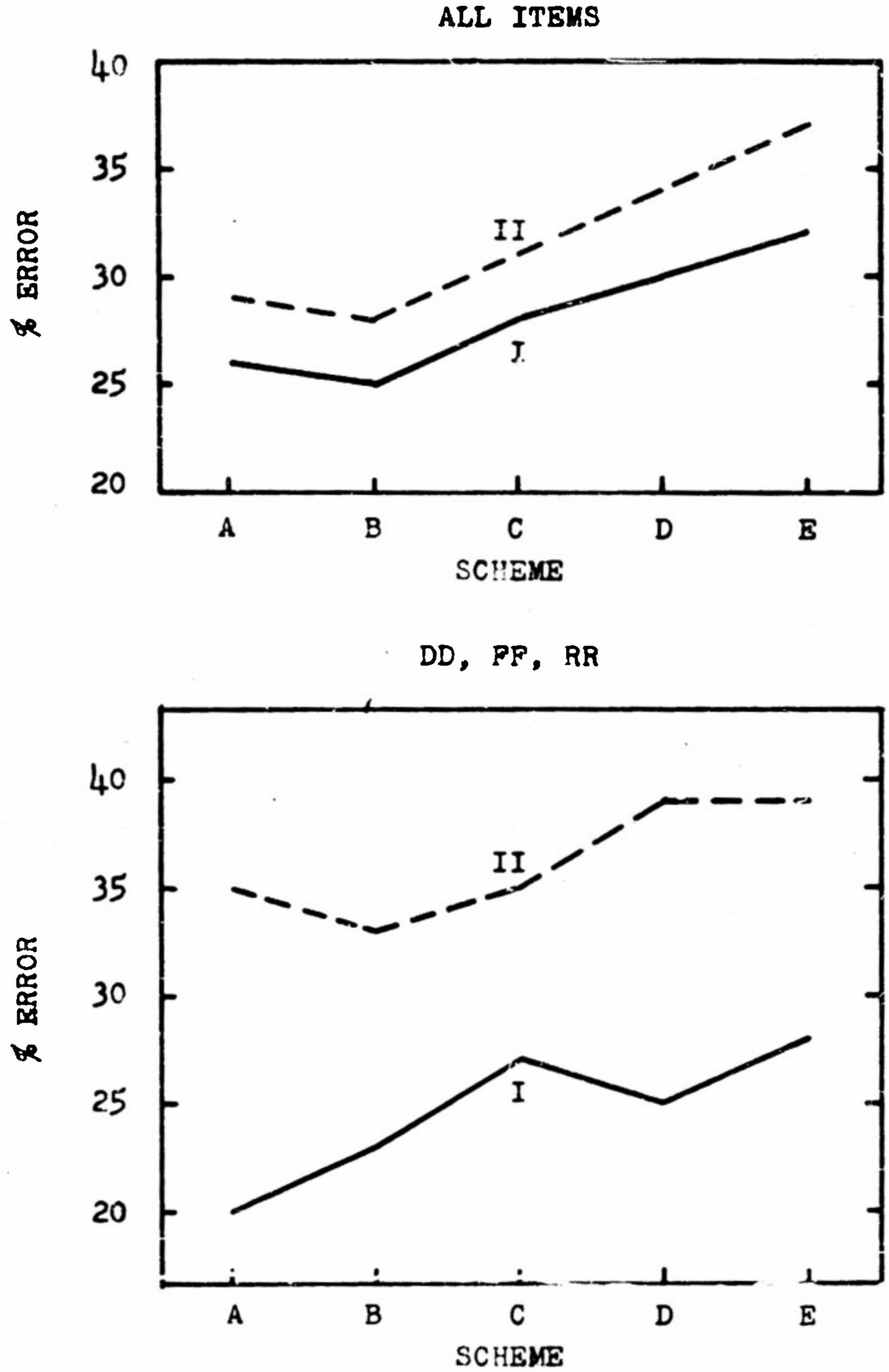


Fig. 5. Average Surface Forecast Errors (errors in %) According to Verification Systems I and II for Schemes A Through E. Wind direction and speed, and precipitation amounts below; all forecast items above.

different forecast items, with the result that the more difficult items such as ww, $h_c h_c$, and $T_d T_d$ tend to dominate thetotal scores. This fact requires further study concerning the forecast s coring system, particularly with respect to the forecast limits. It is also important to note that the forecast sydtem and the scoring system applied to the main project will be somewhat different from those used in this pilot test. For example, VV, ww, RR, and $h_c h_c$ will be forecast in greater detail. Incidentally, the forecasters will remain uninformed of the scoring system to avoid forecast bias.

(5) A more rigorous verification does not greatly influence the trend of the score, although it tends to increase the error. This statement can, obviously, be valid only within certain limits for reasons similar to those mentioned in conclusion 2 above.

The pilot test gave the following results when the upper air data were analyzed:

(1) The difference between 12- and 24-hour forecasts was not negligible for the 12-hour height change (ΔH) and the 12-hour temperature change (ΔT), although the differences between 12- and 24-hour forecasts in wind speed (FF) and direction (DD) was still slight. Figure 6 shows the effect of 12- and 24- hour forecasts on each forecast item for the various schemes. The circled points indicate that in the particular scheme, the item represented by the curve was missing from the analysis information which the forecaster had available.

(2) The average error score for Forecaster I was 28% in the 12-hour forecasts and 34% in the 24-hour forecasts. Forecaster II averaged 24% in 12-hour and 34% in 24-hour forecasts.

(3) The effect of the various schemes on each particular forecast item is not at all clear. However, Figure 6 shows that for the 24-hour forecast of ΔH, the accuracy for the schemes in which the height was missing was considerably less than for schemes in which the height was given.

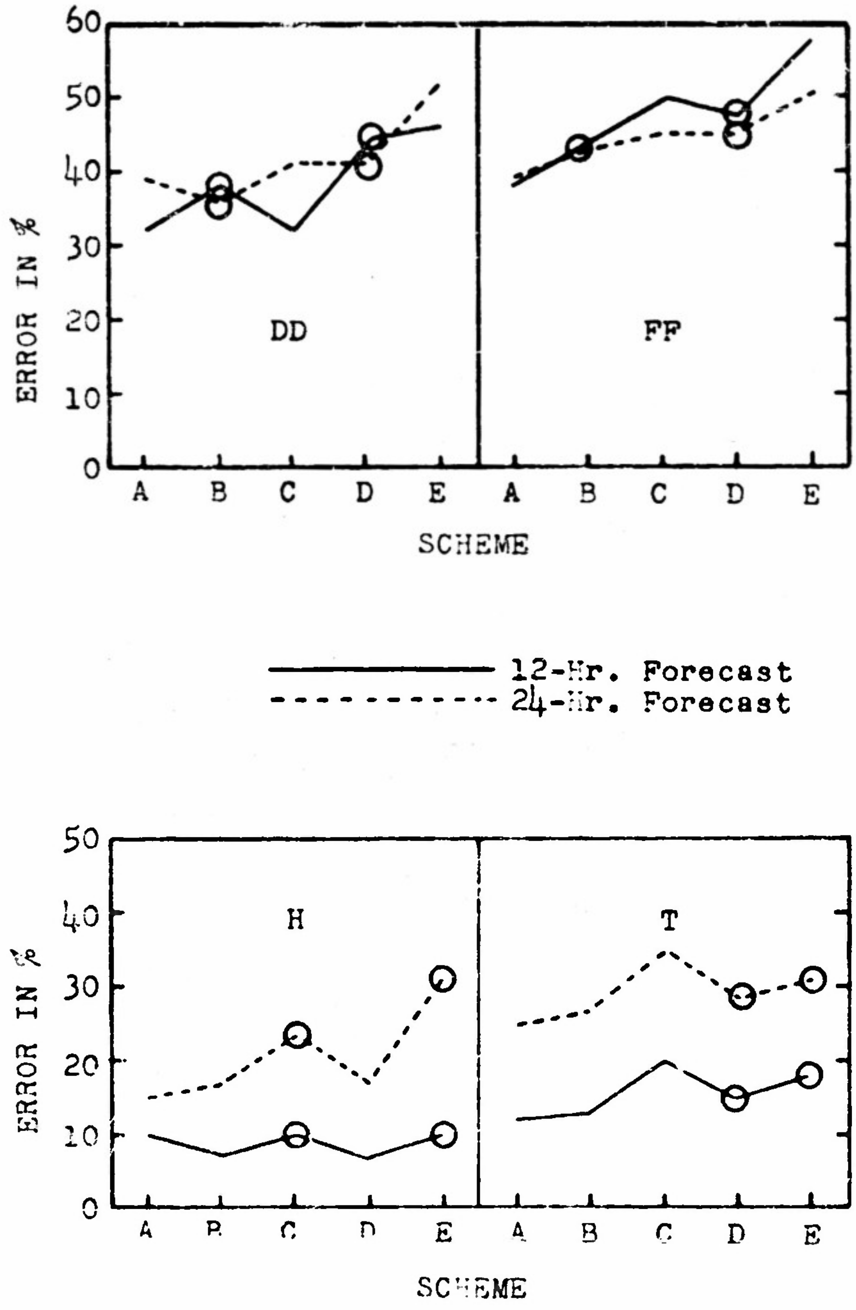


Fig. 6. Effect of Various Schemes on the 500-mb Forecast Items for 12-hr. and 24-hr. Forecasts. Circled Schemes Indicate that the Particular Forecast Item Represented is Missing from that Scheme.

(4) The over-all deterioration of forecast accuracy appears significant only for scheme E. The average per cent error for schemes A and B is 25% and 27%, for schemes C and D it is 31% and 30%, respectively, and for scheme E the average per cent errors reached a maximum of 36%.

Figure 7 combines both 12- and 24-hour forecasts and represents the per cent errors for each station and scheme. No definite deterioration trend is evident. For example, the per cent errors for station 747 seem to decrease with decreased information, whereas stations 304 and 4YE show the opposite trend.

As was done in the case of the surface test, a relative frequency distribution was determined for four groups of per cent errors and for all schemes. The distribution is shown in Figure 8. No pronounced trend seems evident in this distribution.

From the above results the following conclusions are stated and discussed:

(1) There is a pronounced difference between the 12- and 24-hour forecasts of temperature change and height change. Whether this fact is the result of the verification system or of the variability of the forecast parameter in question, is not certain at this time. However, there seems to be no reason why temperature and height should vary to any greater extent than wind direction and speed.

(2) The average scores of the two forecasters were very similar even when divided into 12- and 24-hour forecast categories. This conclusion seems to go along with similar results obtained in the surface pilot test analysis.

(3) Score deterioration as a function of the scheme used does not seem to hold as well as the case in the surface pilot test. Only on the average did scheme E of the upper air test show any marked deterioration. When the data were broken up into individual stations, the deterioration as a function of the scheme showed opposite effects in many cases.

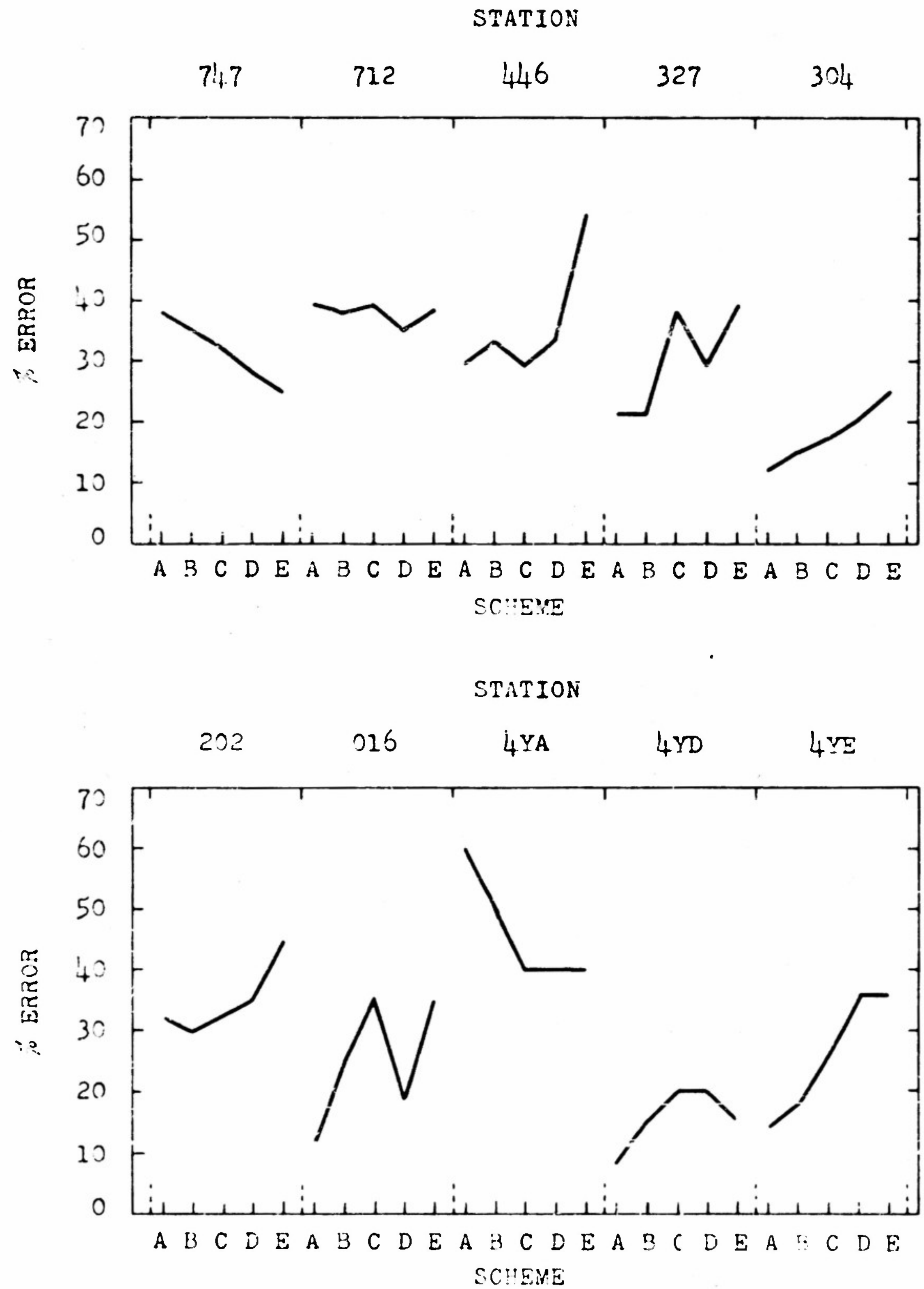


Fig. 7. Average Upper Air Forecast Errors (Errors in %) for Various Stations. Schemes A Through E.

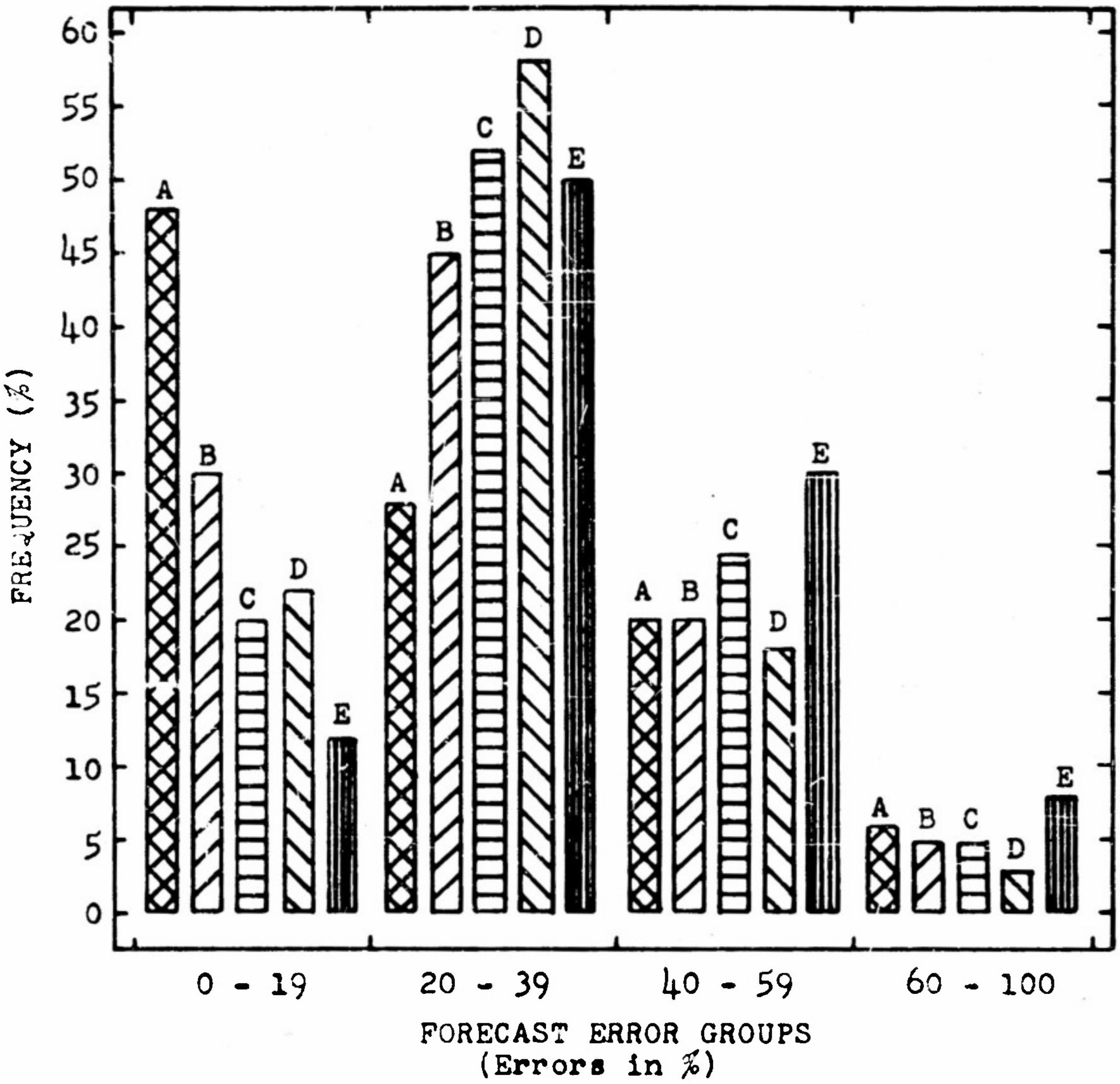


Fig. 8. Frequency Distributions of Upper Air Forecast Errors (Errors in %). Schemes A Through E.

In addition to the score analysis of the surface and upper air pilot tests, a time study was also made. Table V gives the average time in minutes spent in spotting, analyzing, and forecasting for each scheme.

Table V. Average time in minutes spent in the spotting of, analyzing of, and forecasting from the map series.

Scheme	A	B	C	D	E
Spotting	137	102	93	64	35
Analysis	85	94	96	69	84
Forecasting	48	61	61	47	78
Average	79	75	73	52	53

An inspection of the averages listed in Table V shows that schemes D and E took 30% less time than schemes, A, B, and C. Obviously, a major factor in this 30% difference is the decreased time spent in spotting the reduced information. Analysis time remained fairly constant, whereas forecasting time was greatest for scheme E and least for scheme D. This last result is obviously due to the fact that a forced response is required of the forecaster. He naturally objects to, and consequently hesitates in forecasting from a map that he feels furnishes him with insufficient information to perform the task at hand.

In <u>tentatively</u> appraising the optimum over-all conditions for spotting, analyzing, and forecasting with due consideration of the forecast scores, it appears that scheme D represents the amount of information that reduces the time to the greatest extent without significantly affecting the forecast accuracy.

As a result of the pilot test the following general remarks can be made:

(1) One of the primary results of the pilot test was to give the project members a "feel" for the procedure and the difficulties involved. For instance, as a result of going through the five surface schemes, it was decided to add a sixth (F) scheme, since it was thought necessary

to start off the analysis procedure by forcing the forecaster to draw air-flow lines without the aid of pressure data.

(2) The test clearly brought out the fact that station location and sometimes the length of the forecast should be considered as parameters when the main test is analyzed.

(3) The most perplexing difficulty is the choice of the verification limits. This problem is now being thoroughly reviewed by staff members not directly connected with the project. Fortunately, the work in analyzing and forecasting will not be delayed because of this difficulty, since the forecaster will not be informed of the verification limits. In addition, in case two verification systems are decided upon, the punch-card system to be used will allow space for two such verification schemes.

(4) An important part of the test was the time study. In deciding the over-all feasibility of one information scheme against another, the amount of time saved in spotting, analyzing, and forecasting should represent an important factor.

4.2 Prognostic Chart Experiment

On 11 August 1952, Mr. David Jones initiated an experiment in the regular synoptic meteorology laboratory classes in which 37 undergraduate and special (Air Force officer-trainee) students participated. This test consisted of (1) dividing the group into two sections, each of which worked independently without opportunity for intercommunication, (2) giving one section prepared prognostic charts (on the forecast forms) that consisted of the actual positions of fronts and pressure centers as determined by pre-analysis (by project members) of the next two maps (i.e.. 12- and 24-hour "prognostic" positions), (3) requiring the other section to construct its own prognostic charts, and (4) requiring all students to make 12- and 24-hour forecasts for the following stations:

(1) Moosonee, Ontario, 836

(2) International Falls, Minnesota, 747

(3) Caribou, Maine, 712

(4) Chicago (Joliet), Illinois, 536 (JOT)

(5) Buffalo, New York, 528

(6) Nantucket, Massachusetts, 506

(7) Kansas City, Missouri, (Fort Leavenworth, Kansas), 446 (FLV)

(8) Nashville, Tennessee, 327

(9) Hatteras, North Carolina, 304

(10) Brownsville, Texas, 250

(11) Miami, Florida, 202

(12) Kindley AFB, Bermuda, 016

Forecast items included in this test were: sky cover, wind direction and speed, present weather, ceiling, visibility, precipitation amount, temperature, dew point, 700-mb temperature, 700-mb height, 700-mb wind speed and direction. The maps used consisted of a six-map series of North American sea-level and 700-mb charts of 12-hour continuity. They were obtained from Weather Training Supplies, Inc., Cambridge, Massachusetts, and covered the period from 1835Z on 15 December 1946 through 0635Z on 18 December 1946.

A second test was conducted under the same conditions but using Northern Hemisphere sea-level maps of 24-hour continuity, which had been obtained from the Department of Meteorology, New York University for the period from 1300Z on 1 March 1937 through 1300Z on 6 March 1937. Only 24-hour sea-level forecasts for the following stations were required:

(1) El Paso, Texas

(2) Cape Race, Newfoundland

(3) London, England

(4) Algiers, Algeria

(5) Moscow, U. S. S. R.

(6) Bukhara, W. Uzbek S. S. R.

(7) Irkutsk, Irkutsk

(8) Canton, China

(9) Dutch Harbor, Alaska

(10) Honolulu, Hawaii

Forecast items included in this test were sky coverage wind direction and speed, present weather, ceiling, visibility, past weather, and temperature.

The verification systems used in both tests are similar to that used in the pilot test.

4.21 Results of Prognostic Chart Experiment

The 37 students were divided into two groups of 18 and 19, respectively, and placed in different rooms. The groups were instructed to avoid discussion, comparison, or collusion of any kind while analyzing charts and making forecasts. Group A was issued surface prognostic charts for 12 and 24 hours from the time of the maps the students were to analyze. Group B was issued the same pre-plotted charts for analysis but no prognostic charts. Half way through the experiment this procedure was reversed: Group B was issued the prognostic charts, Group A was not. That step was taken to cancel whatever forecasting superiority one group might have over the other.

A total of 444 forecasts, of 13 items each, for 12 stations was obtained from the first (N. American and N. Atlantic) 6-map series; 218 forecasts, of 8 items each, for 10 stations from the second (N. Hemisphere) map series. For the first series the 444 forecasts were composed of four groups of 111 forecasts; 12-hour forecasts with prognosis, 12-hour forecasts without, 24-hour forecasts with, and 24-hour forecasts without. In the second series the 218 forecasts, for which only 24-hour continuity was available, were composed of two groups of 111 with and 107 without prognostic charts. The forecasts were verified on a "percent-correct" basis and mean group scores computed, showing the following results:

First 6-Map Series

Mean score 12-hr. forecast with prog.......65.9%

12-hr. forecast without prog....64.7%

Superiority prog over no-prog 12-hr forecast.....1.2%

Mean score 24-hr forecast with prog.............60.2%

24-hr forecast without prog..........57.7%

Superiority prog over no-prog 24-hr forecast.....2.5%

Second 6-Map Series

Mean score 24-hr forecast with prog............67.2%

24-hr forecast without prog.........68.6%

Superiority no-prog over prog.................. 1.4%

4.211 Statistical Analysis of Results of Prognostic Experiment. Because the differences in scores between the prog and no-prog charts were small, it was necessary to determine whether or not these differences were statistically significant. No significance test was made for the results of the second map series for reasons which are outlined in section 4.213. A frequency distribution for the 444 forecasts of the first map series was compiled and a frequency polygon constructed. The first frequency polygon was revised with seemingly unnecessarily small class intervals for reasons discussed in section 4.212. The frequency polygon is shown in Figure 9. Standard deviations computed by the "short" method were as follows:

Standard Deviations of 12-hour Forecast Scores

with prognostic charts	5.59%
without prognostic charts	6.05%

Standard Deviations of 24-hour Forecast Scores

with prognostic charts	6.82%
without prognostic charts	7.52%

The standard deviation of the differences of means, σ_d, was computed from the formula

$$\sigma_d = \left[\frac{\sigma_a^2}{N_a - 1} + \frac{\sigma_b^2}{N_b - 1} \right]^{1/2}$$

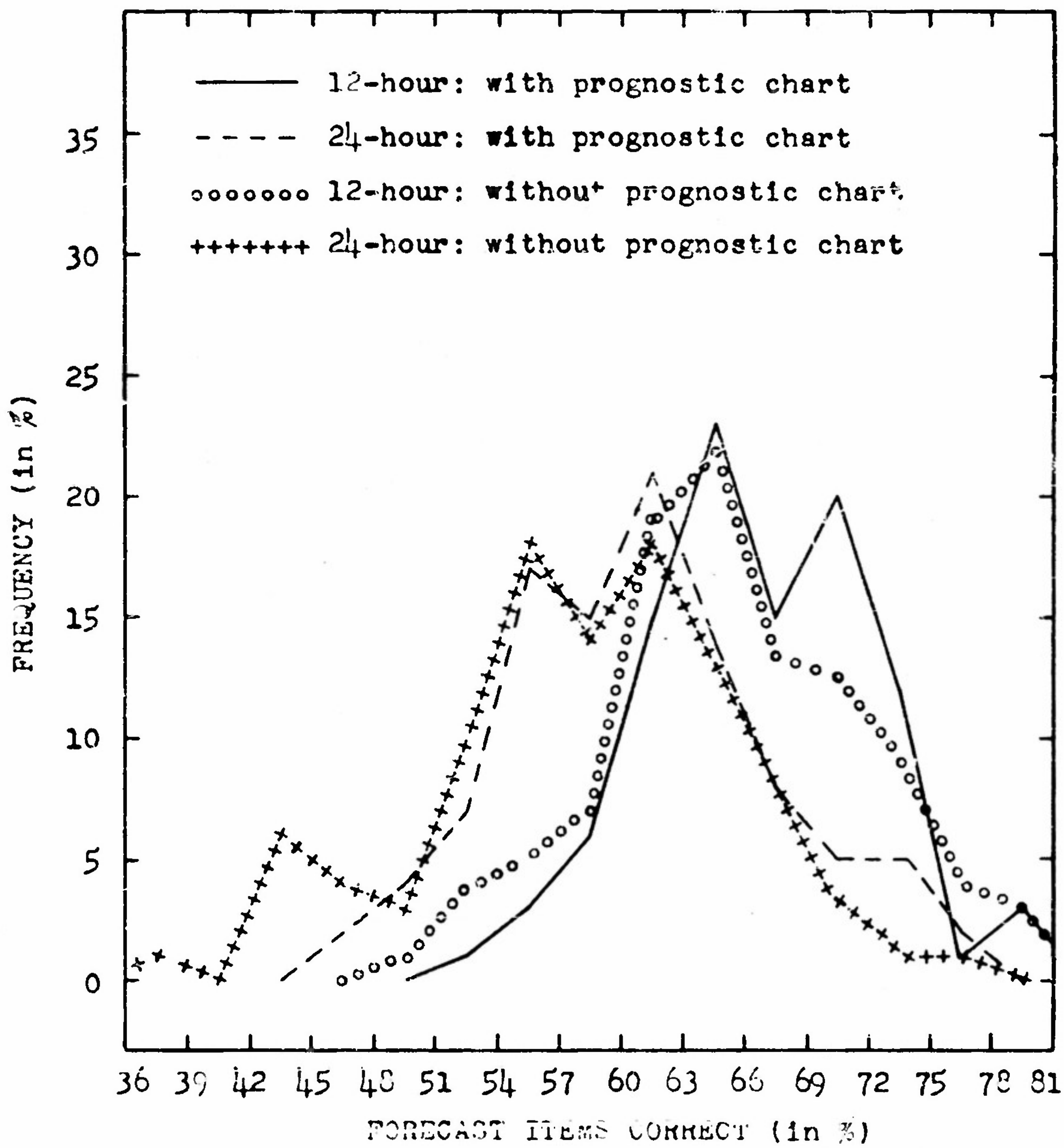


Fig. 9. Frequency Polygon for Unadjusted Forecast Scores. Prognostic Chart Experiment.

where σ_a^2 and σ_b^2 are the variances of forecasting scores in Group A and Group B, respectively, and N_a and N_b the number of participants in Group A and Group B, respectively. The means differed by 1.63 standard deviations of the differences of the means for the 12-hour forecasts and 2.65 standard deviations of the differences of the means for the 24-hour forecasts. Taking, conservatively, 3.0 standard deviations as a basis for determining significance, the percentage differences between the forecasts with prognostic charts and those without were not significant. However, the underlying assumption that the above forecasts were uncorrelated was not justified since a low score by Group A was accompanied by a low score by Group B in the case of difficult forecasts; high scores by both Groups in the case of relatively easy forecasts. That fact led to the development of a more complicated statistical significance test.

4.212 Correction Factor Applied to Statistical Results of Prognostic Chart Experiment. Five factors contributed to the variation of the individual forecasters' scores in the prognostic chart experiment: (1) superior forecasting ability of one Group; (2) presence or absence of prognostic charts; (3) difficulty of the synoptic situation; (4) individual forecasting skill; and (5) random influences. Variations in scores because of (1) were eliminated by computing mean scores for each forecast type and by rotating issuance of prognostic charts between the two Groups. This left (2), (3), (4), and (5) responsible for producing the differences in the mean scores made under each forecast type.

The effect of the difficulty of different forecasts is shown by the following example: The mean score for all participants for the 24-hour forecast from the third map was 50.0%, whereas the mean score for all participants for the 24-hour forecast from the sixth map was 68.6%-- a difference of 18.6%. It was observed that if all six forecasts of each type could be reduced to the same degree of difficulty, the remaining difference could be attributed only to (2), (4), and (5).

The difficulty of a forecast can be measured by the difference between the mean score for all forecasts and the mean score for a particular forecast. If this difference is added to each individual forecaster's score to yield a corrected score, this corrected score should be independent

of the difficulty of the forecast. Furthermore, it seems reasonable to assume that no correlation exists between the corrected scores of Group A and those of Group B. With this assumption, corrected scores were computed and their standard deviations found from the standard deviation of the uncorrected scores by a formula derived by Dr. Hans Panofsky. The derivation of the formula is reproduced and explained in the Appendix to this Progress Report. The standard deviations of the corrected scores were as follows:

Standard Deviations of Corrected 12-Hour Forecast Scores

with prognostic charts................4.16%
without prognostic charts.............4.91%

Standard Deviations of Corrected 24-Hour Forecast Scores

with prognostic charts...............4.24%
without prognostic charts............4.96%

The fact that these values are considerably smaller than the standard deviations of uncorrected forecast scores indicates a great influence of the relative forecast difficulty on the variation of the forecast scores. Quantitatively, the variance because of the variation of the differences of forecast scores accounts for approximately half the total variance of the forecast scores.

From the standard deviations of the corrected forecast scores, σ_d was recomputed. The differences in the means of scores of Groups A and B were 2.16 σ_d for the 12-hour forecasts and 4.01 σ_d for the 24-hour forecasts. According to these figures, the score difference in favor of the prognostic chart's effect on the forecasts is significant in the case of the 24-hour forecasts, and possibly so in the case of the 12-hour forecasts, depending, of course, on whether the 2- or 3-standard deviation limit is used as a criterion of significance. The fact that the differ nce in scores is statistically significant may have little practical value since the difference

in the case of 12-hour forecasts was only 1.4% and in the case of 24-hour forecasts 2.5%.

The frequency of corrected scores in any one class interval of the frequency distribution was greatly increased, so that class intervals of 3%, as in Figure 9, seemed justified. The frequency polygon constructed for the corrected scores is shown in Figure 10.

4.213 Additional Factors Contributing to Results of Prognostic Chart Experiment. The statistical results shown in the preceding sections are necessarily based on the assumption that the forecasts by each man were made independently. Although there was a supervisor in each laboratory at all times throughout the experiment, assurance that there was no collusion between forecasters was not positive; in fact during the second (Northern Hemisphere) series it became clear that the requirements outlined in section 4.21 were not being followed. Several forecasters admitted that they did not use the prognostic chart when making their forecasts. Others found it expedient to retain by memory, or copy, the prognostic chart from day to day. Because of these circumstances it was felt that the results of the second series were meaningless; hence no statistical significance tests were attempted. The laboratory supervisors felt that such was not the case during the first (North American) series, however; hence the results may be indicative of the effect of a "100% accurate" prognostic chart.

Statistical significance in general means that a conclusion is likely to be valid for a population out of which the given sample has been drawn at random. In particular, this means that the experiment described has only shown that prognostic charts are useful for the type of personnel tested, but are not necessarily useful, or may be more useful, for experienced forecasters in the field.

Because of the difficulties involved in supervising a large laboratory group, the experiment is being repeated at present using a small number of men doing a large number of forecasts. Furthermore, it is planned to verify only those forecast items that would definitely be affected by the presence or absence of a correct prognostic chart.

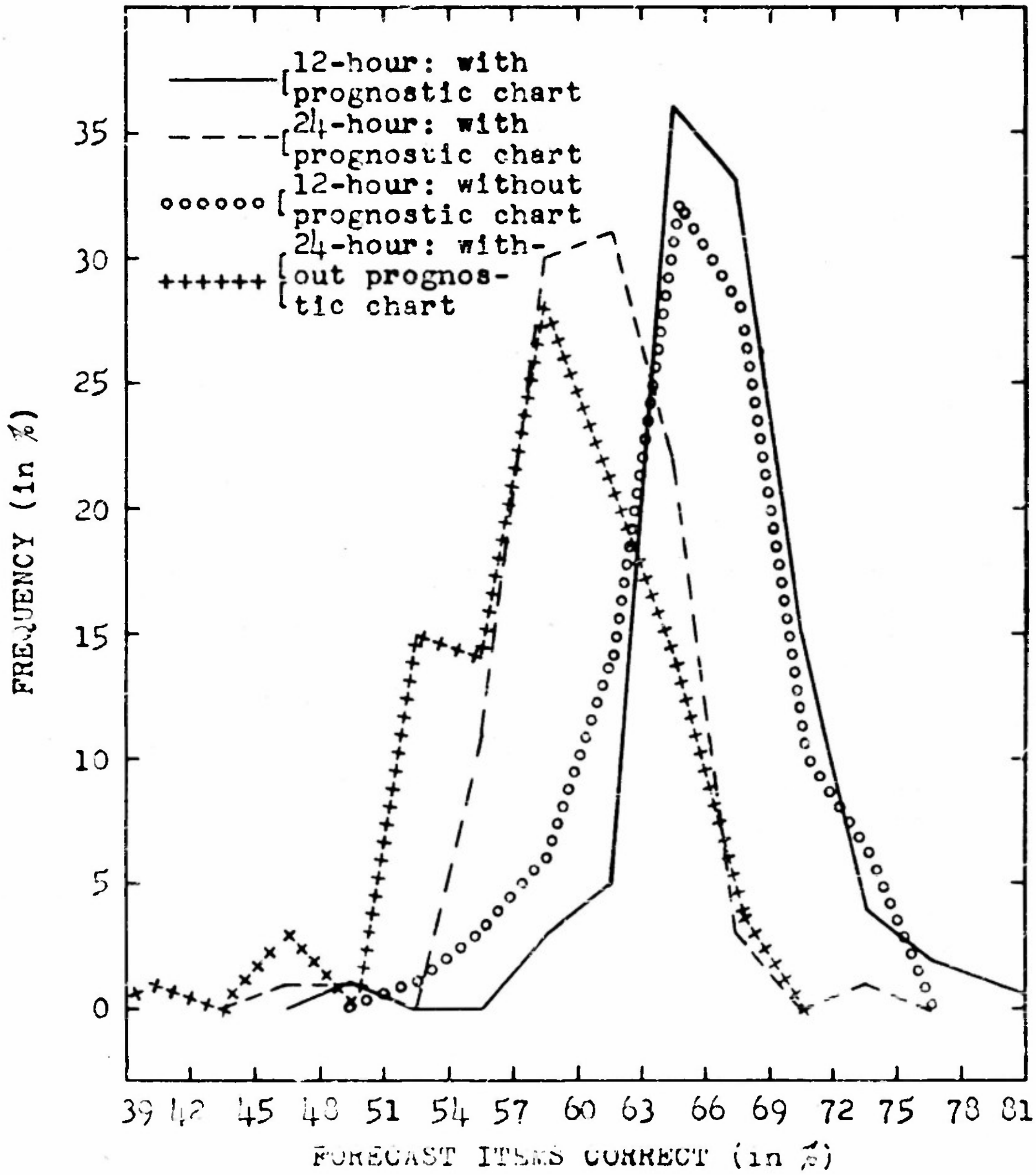


Fig. 10. Frequency Polygon for Adjusted Forecast Scores. Prognostic Chart Experiment.

5.0 Future Plans

(1) The project team is anticipating a visit, early in October by LCDR Paul M. Wolff and LTJG Robert R. Dickson, of Bureau of Aeronautics Project AROWA to State College. The purpose of this visit will be to eliminate evidences of cross-purposes in the investigation and to acquaint LTJG Dickson with the details of previous and planned work in anticipation of his collaboration in the analysis and forecasting of the final test data. LTJG Dickson's participation will be treated on a correspondence basis, in the same manner as that of Mr. Simmons.

6.0 Fiscal Information

July 1, 1952 through September 30,1952

Salaries.	$2,324.00
Wages.	1,132.50
Overhead on Wages and Salaries.	1,420.62
Supplies and Materials	116.22
Equipment.	495.19
Travel.	149.18

7.0 Authorship

The writing of this report was coordinated by the Assistant Project Supervisor, Mr. Moyer, with the active assistance of Mr. Van der Hoven, who wrote up the results of the pilot test, and Mr. Jones, who described the results of the prognostic chart experiment. The drawings were prepared by the two last-mentioned men; and Dr. Panofsky collaborated with Mr. Jones in preparing the Appendix. The editorship of the paper was shared by Dr. Neuberger and Mr. Moyer, and Mrs. Mary Wagner did all the secretarial work as well as the proof-reading.

APPENDIX

Derivation of the Equation From Which the Variance of the Corrected Forecast Scores was Computed.

The variance from an arithmetic mean is defined by the relation

$$\sigma^2 = \frac{\Sigma(X - \overline{X})^2}{N}$$

where X is each participant's score for each forecast, $\overline{X}$ is the mean score for a particular forecast type (12- or 24-hour forecast, with or without prognostic chart), and N is the total number of forecasts in each forecast type (111).

X must be adjusted by adding a correction factor, necessary to reduce all forecasts to the same degree of difficulty. This correction may be designated by $\overline{C}$ and is defined as $\overline{X} - \overline{X}_1$, where $\overline{X}$ is the mean of all forecast scores for each of the 12- and 24-hour groups of six forecasts, and $\overline{X}_1$ is the average for any one of the six forecast scores (scores having the benefit of prognostic charts averaged together with those not having the prognostic charts), one for 12-hour scores, the other for 24-hour scores. Then for each forecast type, the corrected variance, which may be designated by σ_c^2, becomes

$$\sigma_c^2 = \frac{\Sigma[(X + \overline{\overline{C}}) - \overline{X}]^2}{N} \tag{1}$$

or

$$\sigma_c^2 = \frac{\Sigma[(X - \overline{X}) + \overline{\overline{C}}]^2}{N} \tag{2}$$

Expanding the right term of equation (2),

$$\sigma_c^2 = \frac{\Sigma(X - \overline{X})^2}{N} + 2\frac{\Sigma(X - \overline{X})\overline{\overline{C}}}{N} + \frac{\Sigma\overline{\overline{C}}^2}{N} \tag{3}$$

The first term of the right side of equation (3) is the variance of the uncorrected scores which may be designated by σ_u^2. The second and third terms represent the correction to be added to σ_u^2 to arrive at the variance of the corrected forecast scores. The summation sign in the second and third terms essentially represents two successive summations: (1) Σ_i, the sum of men within each laboratory making one forecast; and (2) Σ_j, the sum of all six forecasts either with or without prognostic charts. Σ_i in this experiment is over the six map-days; Σ_j is over the 18 or 19 men in each laboratory.

Rewriting equation (3)

$$\sigma_c^2 = \sigma_u^2 + 2\frac{\Sigma_i\Sigma_j[(X - \bar{X})\bar{\bar{C}}_i]}{N} + \frac{\Sigma_i\Sigma_j\bar{\bar{C}}_i^2}{N} \qquad (4)$$

The forecast difficulty correction factor, $\bar{\bar{C}}_i$, is constant regardless of the number of men in each laboratory. Therefore, it is constant with respect to Σ_j and equation (4) becomes

$$\sigma_c^2 = \sigma_u^2 + 2\frac{\Sigma_i\bar{\bar{C}}_i[\Sigma_j(X - \bar{X})]}{N} + \frac{\Sigma_i\Sigma_j\bar{\bar{C}}_i^2}{N} \qquad (5)$$

The portion of the second term within the brackets in (5) may be expanded: $\Sigma_j X - \Sigma_j \bar{X}$. The term $\Sigma_j X$ is the summation of the average scores for each of the four forecast types for the six map days and may be designated by $m\bar{X}$, where m is either 18 or 19 in this experiment. The term $\bar{X}_i$, the average score of each forecast type for any particular forecast, is defined as

$$\bar{X}_i = \frac{\Sigma_j X}{m} \qquad (6)$$

Rearranging,

$$\Sigma_j X = m\bar{X}_i \qquad (7)$$

The terms $\Sigma_j\bar{X}_i$ and $\Sigma_j X$ (within brackets in equation 5) may be expressed as follows:

$$\Sigma_j(X - \bar{X}) = m\bar{X}_i - m\bar{\bar{X}} = m(\bar{X}_i - \bar{X}) \qquad (8)$$

The difference between the average in each laboratory score for one forecast and the average for the six

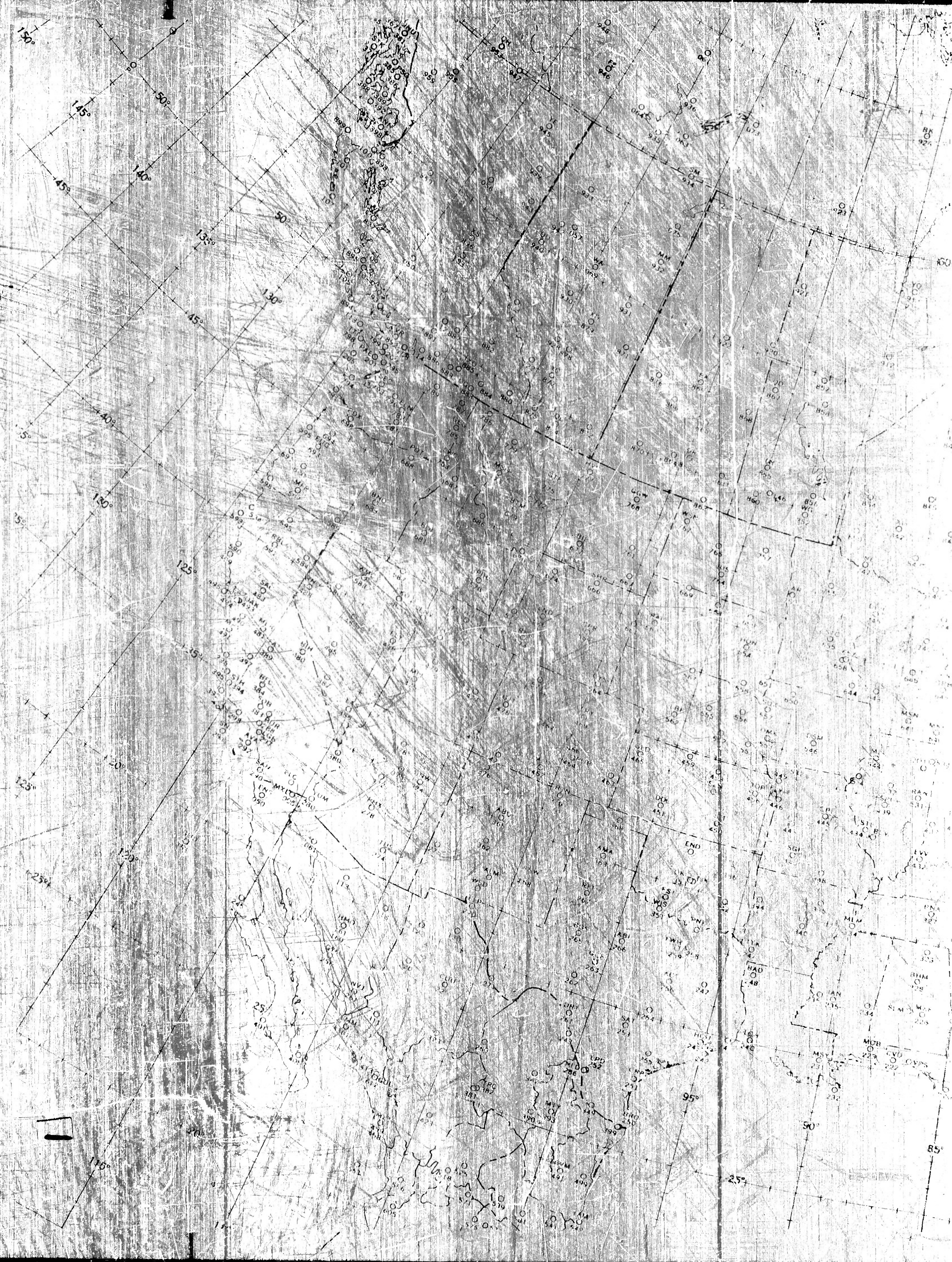
forecasts may be expressed by $\bar{\bar{X}} - X_i$. The difference is constant for each forecast and may be designated as $\bar{\bar{C}}_i$. Therefore,

$$\Sigma_j(X - \bar{X}) = m(\bar{X}_i - \bar{X}) = - m\bar{C}_i \qquad (9)$$

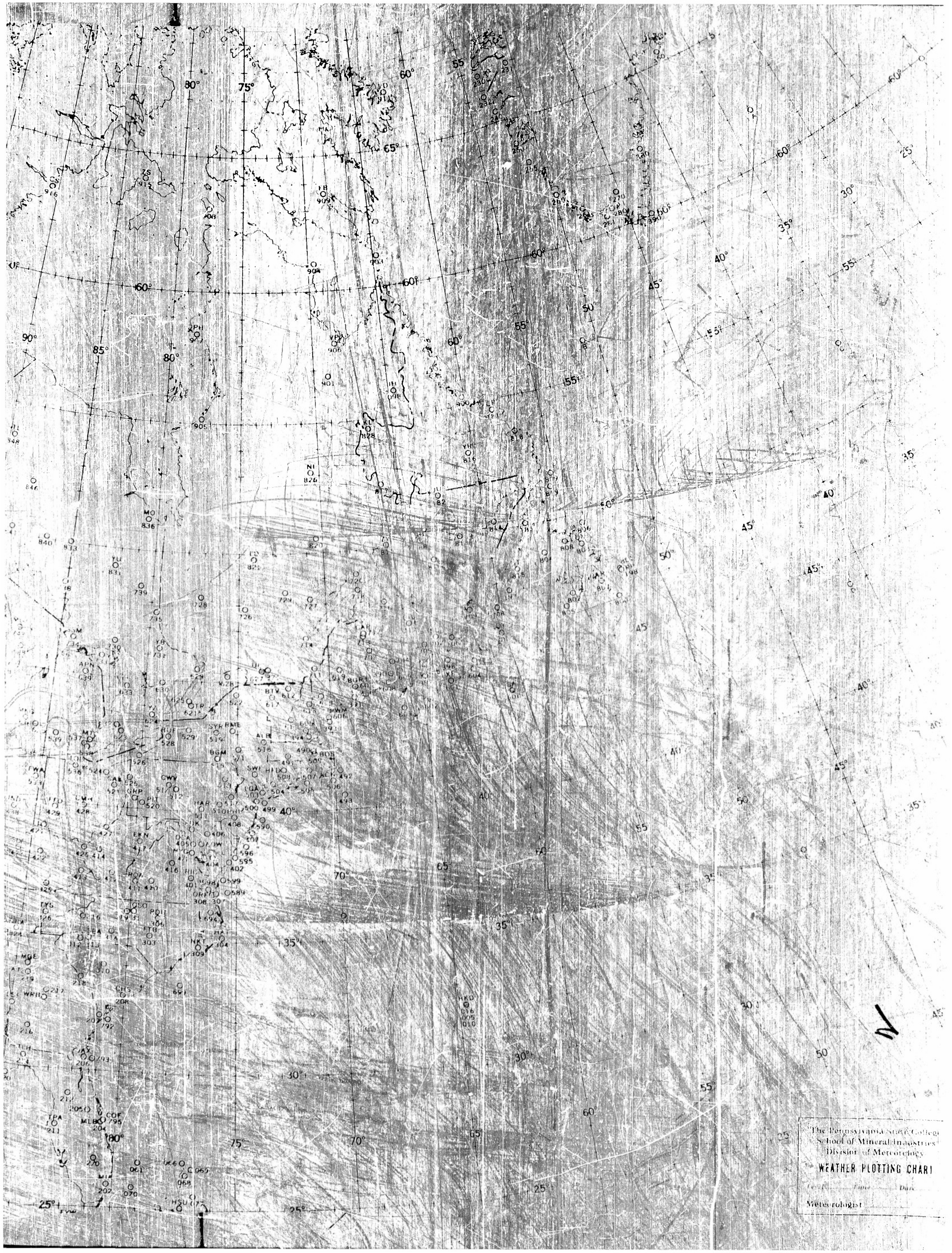
substituting (9) into (5)

$$\sigma_c^2 = \sigma_u^2 - 2\frac{\Sigma_i \bar{\bar{C}}_i m\bar{C}_i}{N} + \frac{\Sigma_i m\bar{\bar{C}}_i{}^2}{N} \qquad (10)$$

With this equation the variance of the corrected scores was computed.



150°
145°
140°
135°
130°
125°
120°
115°
50°
45°
40°
35°
25°
95°
90°
85°


The Pennsylvania State College
School of Mineral Industries
Division of Meteorology
WEATHER PLOTTING CHART
Time
Meteorologist

FORECAST FORM

FORECAST FORM